Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Eighth semester B.Tech degree examinations, September 2020

Course Code: CS466
Course Name: DATA SCIENCE

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks. Marks

1 Write short note on NoSQL. (4)

2 Discuss the need for K–Fold cross validation. (4)

3 (4)

	prediction	
truth	FALSE	TRUE
non-spam	264	14
spam	22	158

Use the above confusion matrix to evaluate various performance metrics of the classifier.

4 (4)

Name	Age	Height	Weight	Sex
Alex	25	177	57	F
Lilly	31	164	69	F
Mark	32	158	56	M
Oliver	26	155	66	M
Martha	28	167	69	F
Lucas	76	172	72	M
Caroline	49	163	66	F

Use the above table to create a data frame; afterwards invert ***sex*** for all individuals. (use R)

5 Create an array with 4 rows and 5 columns and with elements from 1 to 20. Also print the array. (use R) (4)

6 Why area plots is important? Explain how to create an area plot in Python. (4)

7 How hadoop framework helps to meet the big data challenges. (4)

8 With the help of a figure explain Map and Reduce execution. (4)

9 How will you present your model to end users? (4)

10 Discuss deployment methods to demonstrate predictive model operation. (4)

PART B

Answer any two full questions, each carries 9 marks.

11 Illustrate with an example different stages of a data science project. (9)

12 Explain how to evaluate the problems that can be mapped to machine learning techniques. (9)

13 a) Write a note on linear regression. (4)

b) Consider the following data set consisting of the scores of two variables on each of seven individuals: (5)

Subject	A	B
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

Apply K- Mean's algorithm to the above data set to group into two clusters.

PART C

Answer any two full questions, each carries 9 marks.

14 a) Write R codes for constructing, modifying and concatenating lists with examples. (3)

b) The below data set shows the percentage of death rate of women and men in rural and urban region, (6)

Range	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

i. Create the above data frame.

ii. Create a new variable, named ***Total***, which is the sum of each row.

iii. Change the order of the columns so that ***Total*** will be the first variable.

15 With the help of a case study, Explain item based collaborative filtering in Python. (9)

16 a) Construct a scatter plot and bar plot in Python. (4)

b) Discuss statistical models in R. Write two examples. (5)

PART D

Answer any two full questions, each carries 12 marks.

17 a) Describe the various components of Hadoop? (8)

b) Illustrate the master slave architecture work in the Hadoop? (4)

18 a) How to cope with node failures in Hadoop MapReduce? (6)

b) Explain how to do multiple plots in a single window. (6)

19 Discuss the contents of an effective presentation with example. (12)

****